Membrane Protein Structure Predictions for Exploration

Nick V. Grishin[1,*]
[1]Howard Hughes Medical Institute, Departments of Biophysics and Biochemistry, University of Texas Southwestern Medical Center, 6001 Forest Park Road, Dallas, TX, 75390-9050, USA
*Correspondence: grishin@chop.swmed.edu
DOI 10.1016/j.cell.2012.06.004

A daring experiment is performed. Using sequence alignments to predict contacts between residues in protein spatial structures, Hopf et al. are publishing untested de novo structure models for 11 transmembrane protein families. Will their models stand the test of time and hold up to experimentation? The prospects are excellent.

Most proteins function as three-dimensional shapes. These structures are determined by amino acid sequences. The protein folding problem—the accurate prediction of spatial structure from sequence—remains the major challenge of computational biology for more than half a century (Dill et al., 2008). Though faster computing, more accurate energy calculations, and increasingly efficient algorithms to sample protein conformations promise to converge on the ultimate solution, approaches exploiting evolutionary information remain most useful for biologists. In this issue of *Cell*, Hopf et al. (2012) demonstrate just how powerful an evolutionary approach can be for prediction of membrane protein structures. They apply their new method to deduce interresidue contacts from positional correlations in large and diverse sequence alignments.

Proteins are rarely unique. They usually exist in homologous families with similar sequences that have been diversified and polished by evolution to perform needed functions. Because similar sequences typically yield similar spatial structures, an experimentally determined structure for one family member offers reliable structure prediction for the rest of its members. This homology modeling approach is a practical surrogate for the out-of-reach solution of the folding problem (Moult et al., 2011).

What makes homology structure predictions so powerful that they succeed for three out of four proteins? A single sequence contains information for a protein to fold and function. However, as a consequence of evolved tolerance to mutations, this information hides behind noise. An alignment of similar but diverse sequences averages the noise out and reveals common properties, such as functionally important positions and secondary structure. Removal of this noise by alignment is a secret behind success of methods to find very distant homologs with known structure used as templates for modeling (Remmert et al., 2011). These methods assume independence between positions in alignments.

Stepping further, can evolutionary signals in sequences be used for prediction in the absence of experimental structures? Spatial structure is formed by contacts between residues rather than by independent positions. Such contacts have left imprint in the sequences and, if uncovered, would offer structure prediction from alignments without a need for structurally characterized homologs. This simple idea—that correlated mutations should be predictive of contacts—has been applied in many previous studies, including those by Sander and colleagues (Shindyalov et al., 1994). However, these couplings are second-order effects, and thus they are weak. They also form networks of correlated positions so that not all residues with correlated mutations are in direct contact with each other; thus, they were not particularly useful for structure prediction (Fodor and Aldrich, 2004).

Last year, new hopes were raised that correlated mutations could indeed provide key information for structure prediction (Marks et al., 2011). Now, the first practical application of this method has materialized here (Hopf et al., 2012). The trick is to mathematically tease out the correlations caused by direct interactions of residues (i.e., direct coupling) from those caused by contacts through intermediate residues in interaction networks. The main premise is to analyze all of the couplings together instead of individually.

Hopf et al. (2012) show that this prediction strategy is capable of producing excellent 3D models for multihelical membrane proteins. The only input for their software (called EVfold) is a significantly diverse sequence alignment of a protein family. The alignment is used to compute covariation between its positions with the new strategy, which is highly predictive of a contact matrix in the spatial structure. These predicted contacts are used as constraints to generate a structure model most consistent with them. Though not all correlations signify structural proximity of positions, apparently most of them do, which accounts for a good success rate in predictions. The main output of the program is a set of 3D coordinates of a representative structure. Benchmarked on essentially all structurally characterized large families, EVfold gives de novo models without using templates comparable in accuracy to homology models built on distant templates. Such predictions define a general spatial trace of the protein chain but have better details and accuracy around functional sites, which is very important for experimentalists. Additionally and most importantly, Hopf et al. (2012) offer 3D models for

11 transmembrane protein families for which no experimental structures exist. The models are available for the scientific community to test. Even low-resolution structure models bring needed insights to experimental design (Salahudeen et al., 2009). More accurate models help experimental structure determination (Raman et al., 2010).

The predictions resulting from this strategy almost seem too good to be true, given that so many previous studies have already tried to explore positional correlations. What is the trick? I think that it is a combination of innovative theoretical approaches with the avalanche of protein sequences available today. Prediction of direct contacts works well for alignments with a large number of diverse sequences. In other words, this method yields positive results only for protein families with extensive sequence information. Though this dependency on large sequence families boosts confidence in results, it is also a limitation because representative experimental structures are already available for the majority of large protein families. In this case, homology modeling can offer solid predictions instead. So then what is the niche for the new method?

Apparently, transmembrane proteins are a perfect target. Although progress in experimental structure determination of membrane proteins has been steady (Kloppmann et al., 2012), structures of water soluble proteins still outnumber transmembrane proteins by two orders of magnitude. Even for large membrane protein families, representative 3D structures are frequently lacking.

Publishing these unverified de novo structure models is a daring experiment. It has been notoriously difficult to evaluate the state-of-the-art in protein structure prediction. Algorithm development studies usually report ''post-,'' rather than ''predictions,'' benchmarked on proteins with known spatial structures. To offer real prediction opportunities, biannual CASP experiments have been carried out (Moult et al., 2011). An alternative but seemingly more risky approach to record and test truly blind predictions is through their publication. Hopf and colleagues now set precedence in a largely experimental journal by providing this opportunity.

Will their structure models stand the test of time? Will they be useful for biologists working on these proteins? Time will tell, but let's give the models a try. Treat them as low-resolution experimental structures. Design biochemical experiments using them as guides. Use coordinate sets to help determine experimental structures. I predict that most models will be correct, but which ones? Experimentalists will be the judges.

Finally, there is one more lesson to learn. If a large number of diverse sequences having the same 3D structure can predict the structure accurately, perhaps a method can be devised to generate such sequences in cases when they are not available in nature (e.g., for small families). An in vivo selection system for folded proteins could be established; mutations could be introduced and then screened for folded variants. Sequencing the successful variants would then provide large sequence alignments for structure prediction using the method presented by Hopf et al.

ACKNOWLEDGMENTS

This work was supported, in part, by the National Institutes of Health (GM094575 to N.V.G.) and the Welch Foundation (I-1505 to N.V.G.).

REFERENCES

Dill, K.A., Ozkan, S.B., Shell, M.S., and Weikl, T.R. (2008). Annu. Rev. Biophys. *37*, 289–316.

Fodor, A.A., and Aldrich, R.W. (2004). Proteins *56*, 211–221.

Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C., and Marks, D.S. (2012). Cell *149*, this issue, 1488–1502.

Kloppmann, E., Punta, M., and Rost, B. (2012). Curr. Opin. Struct. Biol. Published online May 21, 2012. 10.1016/j.sbi.2012.05.002.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). PLoS ONE *6*, e28766.

Moult, J., Fidelis, K., Kryshtafovych, A., and Tramontano, A. (2011). Proteins *79* (*Suppl 10*), 1–5.

Raman, S., Lange, O.F., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G., Ramelot, T.A., Eletsky, A., Szyperski, T., et al. (2010). Science *327*, 1014–1018.

Remmert, M., Biegert, A., Hauser, A., and Soding, J. (2011). Nat. Methods *9*, 173–175.

Salahudeen, A.A., Thompson, J.W., Ruiz, J.C., Ma, H.W., Kinch, L.N., Li, Q., Grishin, N.V., and Bruick, R.K. (2009). Science *326*, 722–726.

Shindyalov, I.N., Kolchanov, N.A., and Sander, C. (1994). Protein Eng. *7*, 349–358.